Published in final edited form as:

Stat Med. 2012 December 20; 31(29): . doi:10.1002/sim.5433.

Variable selection for covariate-adjusted semiparametric inference in randomized clinical trials

Shuai Yuan, Hao Helen Zhang, and Marie Davidian*

Department of Statistics, North Carolina State University, Raleigh, NC, USA

Abstract

Extensive baseline covariate information is routinely collected on participants in randomized clinical trials, and it is well-recognized that a proper covariate-adjusted analysis can improve the efficiency of inference on the treatment effect. However, such covariate adjustment has engendered considerable controversy, as *post hoc* selection of covariates may involve subjectivity and lead to biased inference, while prior specification of the adjustment may exclude important variables from consideration. Accordingly, how to select covariates objectively to gain maximal efficiency is of broad interest. We propose and study the use of modern variable selection methods for this purpose in the context of a semiparametric framework, under which variable selection in modeling the relationship between outcome and covariates is separated from estimation of the treatment effect, circumventing the potential for selection bias associated with standard analysis of covariance methods. We demonstrate that such objective variable selection techniques combined with this framework can identify key variables and lead to unbiased and efficient inference on the treatment effect. A critical issue in finite samples is validity of estimators of uncertainty, such as standard errors and confidence intervals for the treatment effect. We propose an approach to estimation of sampling variation of estimated treatment effect and show its superior performance relative to that of existing methods.

Keywords

covariate adjustment; false selection rate control; oracle property; semiparametric treatment effect estimation; shrinkage methods; variable selection

1. Introduction

The primary goal of a randomized clinical trial is to compare two or more treatments on the basis of some measure of treatment effect. Standard analyses focus on comparisons averaged across the patient population and use data on outcome and treatment assignment only. Often, however, considerable baseline auxiliary covariate information is also collected, and it is well known that precision of inference on the treatment effect may be improved by “adjusting” for relationships between outcome and covariates, discussed in an extensive literature; e.g., [1, 2, 3, 4].

A fundamental debate surrounding such analyses involves variable selection; i.e., how to identify covariates that should be used for adjustment. Concern has been expressed by trialists regarding adjustment based on analysis of covariance (ANCOVA) models for the relationship between outcome, treatment assignment, and covariates, as these link estimation

of the treatment effect to that of covariate effects in a single model and thus invite the temptation for a *post hoc* search for models and variables that best favor the treatment effect. This has led to regulatory authorities to dictate that such analyses be based on prespecified models involving only a few covariates [5], which holds the potential for missed opportunities to detect important treatment effects.

Tsiatis *et al.* [6] and Zhang *et al.* [7] exploited semiparametric theory (e.g., [8]) to develop methods for covariate adjustment under unrestrictive assumptions that involve modeling the outcome-covariate relationship separately by treatment group, thus decoupling covariate effects from estimation of the treatment effect; see also [9, 10, 11].

They argued that this feature supports a strategy where distinct analysts develop treatment-specific models for the regression of outcome on covariates, deploying variable selection techniques to yield the most parsimonious and predictive models, with inference on the treatment effect constructed subsequently and independently. According to the theory in [7], this could lead to maximal efficiency gains over the standard unadjusted analysis.

Despite the promise of this approach, there has been little work on whether or not and how variable selection methods should be used to develop the treatment-specific models in this context. In this article, we study integration of modern variable selection methods with the semiparametric methods of [7], document their performance under a host of conditions, and make recommendations for practice.

Tsiatis *et al.* [6] observed that use of the standard asymptotic “sandwich” method to estimate the true sampling variance of the semiparametric treatment effect estimator could lead to under-estimation of uncertainty and optimistic confidence intervals and speculated that this could be due to finite sample size and the effects of model fitting and variable selection. In the case of estimation of the difference of two treatment means, they proposed a “correction factor” to account for these phenomena but gave no theoretical basis, and generalization to other settings is not apparent. We propose an alternative estimator for sampling variance and demonstrate its promising performance in small and moderate samples.

In Section 2, we review the semiparametric covariate adjustment framework of [7]. In Section 3, we propose several variable selection methods for developing treatment-specific models to be integrated with the semiparametric methods and provide a generic algorithm for implementation. We examine deficiencies of the standard sandwich estimator for sampling variance and propose an alternative estimator and a rationale for why it offers improved performance in Section 4. We report on simulations to evaluate performance of the strategies we propose in Section 5, and we apply the methods to data from a HIV clinical trial in Section 6. We make recommendations for practice in Section 7.

2. Semiparametric Framework and Inference

Consider a K -arm randomized trial with $K \geq 2$. Represent the data from subject i as (Y_i, X_i, Z_i) , $i = 1, \dots, n$, independent and identically distributed (iid) across i , where Y_i is the outcome of interest; X_i is the $(d \times 1)$ vector of all baseline covariates; and $Z_i = k$, $k = 1, \dots, K$, is the indicator of assignment to treatment k , carried out according to known randomization

probabilities $\pi_k = \text{pr}(Z = k)$ with $\sum_{k=1}^K \pi_k = 1$. Randomization guarantees that Z is independent of X , written as $Z \perp\!\!\!\perp X$.

Following [7], let $\beta = (\beta_1, \dots, \beta_K)^T$ be a vector of parameters associated with treatment comparisons under a statistical model describing features of the conditional distribution of Y given Z . When Y is continuous, a standard model would be

$$E(Y|Z) = \beta_1 + \beta_2 I(Z=2) + \cdots + \beta_K I(Z=K), \quad \beta_1 = E(Y|Z=1), \quad (1)$$

so that $\beta_k = E(Y|Z=k) - E(Y|Z=1)$ for $k=2, \dots, K$, and treatment 1 is the reference. For binary $Y=0, 1$, with $Y=1$ indicating the event of interest, a popular model is

$$\text{logit}\{E(Y|Z)\} = \text{logit}\{\text{pr}(Y=1|Z)\} = \beta_1 + \beta_2 I(Z=2) + \cdots + \beta_K I(Z=K), \quad (2)$$

where $\text{logit}(p) = \log\{p/(1-p)\}$; β_1 is the log odds of treatment 1; and $\beta_k, k=2, \dots, K$, is the log odds ratio for treatment k relative to the reference treatment 1. For longitudinal studies, with Y_i a vector of outcomes $Y_{ij}, j=1, \dots, m_i$, at times $t_{i1}, t_{i2}, \dots, t_{im_i}$, one might postulate

$$Y_{ij} = \tau_0 + \{\beta_1 I(Z=1) + \cdots + \beta_K I(Z=K)\} t_{ij} + (b_{i0} + b_{i1} t_{ij}) + e_{ij}, \quad (3)$$

a linear mixed model with common intercept τ_0 ; iid $N(0, D)$ random effects $b_i = (b_{i0}, b_{i1})$ for (2×2) covariance matrix D ; and e_{ij} iid $N(0, \sigma_e^2)$ for all i, j . In (3), $\beta_k - \beta_{k'}$ is the difference in mean slopes for treatments k and k' .

When $K=2$, alternative notation $Z=0, 1$, with randomization probabilities $(1-\pi)$ and π corresponding to treatments 0 and 1 (e.g., control or experimental treatment), is standard. We adopt this without comment in the sequel. Here, writing $E(Y|Z) = \beta_0 + \beta_1 Z$ or $\text{logit}\{E(Y|Z)\} = \beta_0 + \beta_1 Z$, $\beta_1 = E(Y|Z=1) - E(Y|Z=0)$ for general outcomes or $\beta_1 = \text{logit}\{E(Y|Z=1)\} - \text{logit}\{E(Y|Z=0)\}$ for binary outcome is of interest.

Many such models, which are assumed correct for the primary analysis, involve β alone and minimal assumptions on the distribution of (Y, Z) , while others like (3) depend on additional

parameters γ , say; e.g., $\gamma = \{\tau_0, \sigma_e^2, \text{vech}(D)^T\}$ where $\text{vech}(D)$ is the vector of all the distinct parameters in D , and make parametric distributional assumptions. In general, write $\theta = (\beta^T, \gamma^T)^T$ ($r \times 1$), with true value $\theta^0 = (\beta^0{}^T, \gamma^0{}^T)^T$. The standard unadjusted analysis estimates θ , and hence β , based on the data $(Y_i, Z_i), i=1, \dots, n$. In models like (1)–(3), consistent and asymptotically normal estimators for $\theta(\beta)$ may be obtained by solving $(r \times 1)$ unbiased

estimating equations [12, Section A.6] of the form $\sum_{i=1}^n m(Y_i, Z_i, \theta) = 0$. E.g., the ordinary least squares (OLS) estimator for β_1 in (1) takes the estimating function $m(Y, Z, \theta) = [1, I(Z=2), \dots, I(Z=K)]^T \{Y - \beta_1 - \beta_2 I(Z=2) - \cdots - \beta_K I(Z=K)\}$; for $K=2$, the resulting estimator for $\beta = E(Y|Z=1) - E(Y|Z=0)$ is the difference in sample means $\bar{Y}^{(1)} - \bar{Y}^{(0)}$, where $n_i = \sum_{i=1}^n Z_i, n_0 = n - n_1, \bar{Y}^{(1)} = n_1^{-1} \sum_{i=1}^n Y_i Z_i$, and $\bar{Y}^{(0)} = n_0^{-1} \sum_{i=1}^n Y_i (1 - Z_i)$. Similarly the maximum likelihood estimator for β in (2), and that for θ in (3) solve unbiased estimating equations; see [7] for the form of the estimating functions $m(Y, Z, \theta)$.

In a covariate-adjusted analysis, the goal is to make inference on θ and hence β using the data $(Y_i, X_i, Z_i), i=1, \dots, n$. Under conditions given in [7], the authors show all unbiased estimating functions for θ can be written as

$$m^*(Y, X, Z; \theta) = m(Y, Z, \theta) - \sum_{k=1}^K \{I(Z=k) - \pi_k\} a_k(X), \quad (4)$$

where $m(Y, Z, \theta)$ is a fixed unbiased estimating function for θ based on (Y, Z) only; $a_k(X), k=1, \dots, K$, are arbitrary r -dimensional functions of X ; and $\sum_{k=1}^K \pi_k a_k(X) = 0$ ensures that the estimating function $m^*(Y, X, Z; \theta)$ has mean zero. The second term in (4) “augments” the estimating function for the unadjusted analysis to gain efficiency by exploiting the

covariates. Zhang *et al.* [7] showed that the optimal choice of $a_k(X)$, leading to the estimator for $\theta(\beta)$ solving estimating equations of the form $\sum_{i=1}^n m^*(Y_i, X_i Z_i; \theta) = 0$ with smallest asymptotic variance among all such estimators, is

$$a_k^{opt}(X) = E\{m(Y, Z, \theta) | X, Z=k\}, k=1, \dots, K.$$

Common estimating functions in models like (1)–(3), with the covariance parameters known in the latter, may be written for some $A(Z, \theta)$ with r rows and some $f(Z, \theta)$ as

$$m(Y, Z, \theta) = A(Z, \theta) \{Y - f(Z, \theta)\}, \quad (5)$$

and the optimal augmented estimating equations for covariate adjustment become

$$\sum_{i=1}^n \left[m(Y_i, Z_i, \theta) - \sum_{k=1}^K \{I(Z_i=k) - \pi_k\} A(k, \theta) \{E(Y_i | X_i, Z_i=k) - f(k; \theta)\} \right] = 0. \quad (6)$$

In what follows, we restrict attention to unadjusted estimating functions of type (5). In the special case of (1) and $K=2$, so that $\beta_1 = E(Y|Z=1) - E(Y|Z=0)$ with $\pi = \text{pr}(Z=1)$, (6) may be solved to yield the explicit form of the optimal adjusted estimator for β_1 ; although π is known, it is preferable on grounds of efficiency to substitute the estimator $n_1/n = \bar{Z}$ for π in this expression, which yields the estimator for β_1 given by [6],

$$(\bar{Y}^{(1)} - \bar{Y}^{(0)}) - \sum_{i=1}^n (Z_i - \bar{Z}) \{n_0^{-1} E(Y_i | X_i, Z_i=0) + n_1^{-1} E(Y_i | X_i, Z_i=1)\}. \quad (7)$$

From (6) and (7), implementation of the adjustment reduces to modeling the unknown treatment-specific regressions $E(Y|X, Z=k)$, $k=1, \dots, K$. For $k=1, \dots, K$, [6, 7] proposed positing parametric models of the form $q_k^*(X, \xi_k^*)$, say, for $E(Y|X, Z=k)$ for each k ; obtaining estimates $\hat{\xi}_k^*$ via an appropriate method, e.g., OLS or iteratively reweighted least squares (IRWLS) carried out separately within each group k ; and replacing $E(Y_i | X_i, Z_i=k)$ in (6) or (7) by $q_k^*(X_i, \hat{\xi}_k^*)$. In the particular case of (7), this yields the estimator

$$\hat{\beta}_1 = (\bar{Y}^{(1)} - \bar{Y}^{(0)}) - \sum_{i=1}^n (Z_i - \bar{Z}) \{n_0^{-1} q_0^*(X, \hat{\xi}_0^*) + n_1^{-1} q_1^*(X, \hat{\xi}_1^*)\}. \quad (8)$$

As discussed in Section 4 of [7], under regularity conditions, for $q_k^*(X, \xi_k^*)$ linear in ξ_k^* , the resulting estimator $\hat{\theta}$ and hence $\hat{\beta}$ will be consistent and asymptotically normal regardless of whether or not the postulated models are correct. Moreover, if $q_k^*(X, \xi_k^*)$ are fitted by OLS, $\hat{\theta}$ will have the same asymptotic distribution as if the limits in probability of the $\hat{\xi}_k^*$ were known and substituted and is guaranteed to be relatively more efficient than the unadjusted estimator solving $\sum_{i=1}^n m(Y_i, Z_i, \theta) = 0$. Although the theory does not extend strictly to nonlinear $q_k^*(X, \xi_k^*)$, empirical evidence has shown that these properties hold approximately; e.g., for logistic models fitted by IRWLS. If the same model is postulated for each k and fitted using all the data aggregated across groups, similar properties hold; however, the resulting estimator for treatment effect will not achieve maximal efficiency if the true regression relationships differ. In general, if the posited models contain the true $E(Y|X, Z=k)$, the estimator will be asymptotically equivalent to the optimal estimator solving (6). As in (8), estimators for the π_k should be substituted. The estimating equation (6) is unbiased and

hence resulting estimators such as (8) are consistent regardless of whether or not the π_k are the same for all k (i.e., equal randomization).

Zhang *et al.* [7] proposed using routine M-estimator results to obtain estimated standard errors for such $\hat{\beta}$. In particular, as we discuss further in Section 4, they suggested using the square roots of the diagonal elements of the usual sandwich estimator for the large sample covariance matrix of all parameters. In the case $K = 2$ and (1), this yields the estimated sampling variance for the estimator $\hat{\beta}_1$ for $\beta_1 = E(Y|Z = 1) - E(Y|Z = 0)$ in (8) given by [6],

$$\sum_{i=1}^n \left[\left\{ n_1^{-1} Z_i - n_0^{-1} (1 - Z_i) \right\} Y_i - n^{-1} \hat{\beta}_1 - (Z_i - \bar{Z}) \left\{ n_0^{-1} q_0^* (X_i, \hat{\xi}_0^*) + n_1^{-1} q_1^* (X_i, \hat{\xi}_1^*) \right\} \right. \\ \left. - (Z_i - \bar{Z}) \left\{ n_0^{-1} (\bar{Y}^{(0)} - \bar{q}_0^*) + n_1^{-1} (\bar{Y}^{(1)} - \bar{q}_1^*) \right\} \right]^2, \quad (9)$$

where for $k = 0, 1$, $\bar{q}_k^* = n_k^{-1} \sum_{i=1}^n I(Z_i = k) q_k^*(X_i, \hat{\xi}_k^*)$. The final term in (9) is a consequence of substitution of the sample proportion n_1/n for π and is identically equal to zero if the $q_k^*(X_i, \hat{\xi}_k^*)$ are linear models with intercepts and fitted separately by OLS. Tsiatis *et al.* [6] noted that, when n is not large, (9) may understate the true sampling variation and speculated that this is due to second order effects of estimation of the ξ_k^* ; as noted above, first order theory implies no such effect. They proposed that (9) be multiplied by a small sample “correction factor” 1 depending on n_0 , n_1 , and the numbers of non-zero non-intercept terms in each ξ_k^* that approaches one with increasing sample size. However, this has no theoretical justification and does not extend readily to other settings.

3. Variable Selection

The foregoing approach to covariate adjustment, which separates positing and fitting of the $q_k^*(X_i, \hat{\xi}_k^*)$ from final estimation of the treatment effect parameters β , allows objective development of these models via any techniques at the analyst’s disposal. Under the premise that models that get as close as possible to the true $E(Y|X, Z = k)$ will yield maximal efficiency gains over an unadjusted analysis, we propose a strategy, which could be specified in advance in a study protocol, in which modern variable selection methods are used for this purpose. In comparison to traditional techniques such as forward or backward selection with *ad hoc* “ α to enter/leave” criteria, recent approaches that explicitly seek to minimize prediction error and/or improve tuning of traditional procedures may be more suitable.

For definiteness, consider the following framework for variable selection in our context. Suppose one is willing to assume that the form of the true treatment-specific regressions is

$$E(Y|X, Z = k) = q_k(X, \xi_k) = g_k \left\{ \xi_{k0} + \xi_{ks}^T c_k(X) \right\}, \quad k = 1, \dots, K, \quad (10)$$

where $c_k(X) = \{c_{k1}(X), \dots, c_{ksk}(X)\}^T$ is a s_k -vector of functions of X , $\xi_{ks} = (\xi_{k1}, \dots, \xi_{ksk})^T$ is a vector of corresponding coefficients, ξ_{k0} is an intercept term, $\xi_k = (\xi_{k0}, \xi_{ks}^T)^T$, and g_k is a known one-to-one inverse link function. The $c_{kj}(X)$ may include polynomial terms, interactions, and splines in the elements of X . E.g., if the predictor $\xi_{ks}^T c_k(X)$ in (10) includes all linear, quadratic, and two-way interaction terms in elements of X , $s_k = 2d + d(d-1)/2$; $c_k(X) = X$ yields a representation with only linear terms, and $s_k = d$. Assume also that, for

each treatment, if (10) were correct, not all terms in the predictor $\xi_{ks}^T c_k(X)$ are important; i.e., some elements of ξ_{ks} in (10) are zero. Thus, there is an associated important index set, $M_k = \{j : \xi_{kj} \neq 0, j = 1, \dots, s_k\}$ for treatment k , which provides a parsimonious representation of the form $E(Y|X, Z = k) = q_k(X, \xi_k) = g_k\{\xi_{k0} + \sum_{j \in M_k} c_{kj}(X)\xi_{kj}\}$.

Assuming (10), the goal is to estimate the index sets $M_k, k = 1, \dots, K$, correctly from the treatment-specific data. We propose both shrinkage type estimators and forward selection with false selection rate (FSR) control [13]. Assume for simplicity that X is standardized to have mean zero and norm one.

For treatment k , the shrinkage estimators solve

$$\min_{\xi_k} \left\{ L_{n_k}(X, \xi_k) + \sum_{j=1}^{s_k} \psi_{\lambda_j}(|\xi_{kj}|) \right\}, \quad (11)$$

where $L_{n_k}(X, \xi_k)$ is the likelihood or loss function in ξ_k relevant to fitting (10) based on the n_k observations under treatment k ; $\psi_{\lambda}(\cdot)$ is the penalty function; and $\lambda_j, j = 1, \dots, s_k$, are tuning parameters controlling the balance between fit and model size. Various choices of $\psi_{\lambda}(|u|)$ in (11) correspond to popular methods. The Least Absolute Shrinkage and Selection Operator (LASSO) [14] uses $\psi_{\lambda}(|u|) = \lambda|u|$; the Smoothly Clipped Absolute Deviation (SCAD) [15] penalty function has derivative

$$\psi'_{\lambda}(|u|) = \lambda \left[I(|u| \leq \lambda) + (a\lambda - |u|)_+ I(|u| > \lambda) / \{(a-1)\lambda\} \right] \text{ with } a = 3.7 \text{ as}$$

recommended by [15]; and the adaptive LASSO [16] uses $\psi_{\lambda}(|u|) = \lambda|u|/|\tilde{u}|$, where $\tilde{u}$ is a consistent initial estimator of u . Variations include the one-step sparse estimator (OSE) [17], which uses a local linear approximation for the penalty function, and the least squares approximation (LSA) [18] estimator, based on a quadratic approximation to $L_{n_k}(X, \xi_k)$ with the adaptive LASSO penalty. The λ_j are chosen via cross validation or criteria such as the Bayesian Information Criterion (BIC). Alternatively, Boos *et al.* [13] proposed to improve classical forward selection by automatic estimation of the tuning parameter “ α to enter” to achieve a target FSR, the expected proportion of unimportant terms identified out of the total. The method estimates the largest α such that an estimate of achieved FSR is not more than the target, yielding a rule for identifying M_k .

We propose to integrate any of these variable selection procedures with the covariate adjustment framework of the previous section in the following generic objective algorithm:

Step 1

For each treatment $k = 1, \dots, K$, assume a form $q_k(X, \xi_k)$ for the true $E(Y|X, Z = k)$ as in (10). Use a prespecified variable selection method from among those above with prespecified procedure for tuning to estimate M_k . For shrinkage methods, this would be the particular implementation of cross validation or information criterion to be used; for the FSR method, this involves specifying the target FSR. The resulting $\hat{\xi}_k$ will include zero entries and imply estimates $\widehat{M}_k = \{j : \hat{\xi}_{kj} \neq 0, j = 1, \dots, s_k\}$.

Step 2

For each k , let $\hat{\xi}_k^*$ denote the collapsed vector of nonzero entries of $\hat{\xi}_k$, and let $q_k^*(X, \xi_k^*)$ denote the form of assumed model with all estimated zero terms eliminated. Form predicted values $q_k^*(X_i, \hat{\xi}_k^*)$ for all $i = 1, \dots, n$ and $k = 1, \dots, K$.

Step 3

Substitute the $q_k^*(X_i, \hat{\xi}_k^*)$ for $E(Y_i|X_i, Z_i = k)$ in (6) and solve for θ . Denote the final estimate of the treatment effect parameters as $\hat{\beta}$.

Although the performance of all the foregoing variable selection approaches in terms of minimizing prediction error and correctly identifying M_k has been widely studied, little is known about their relative merits in the context of the above algorithm. Here, there is little interest in the treatment-specific models so identified except to the extent that they lead to consistent and efficient estimation of β , so it is not clear whether or not these usual criteria for evaluating variable selection methods are relevant. When the inverse link function g_k is the identity and the model in (10) is correctly specified, so that the true $E(Y|X, Z = k)$ is a linear regression and $L_{n_k}(X, \xi_k)$ is the OLS loss function, if the tuning parameters in (11) are chosen properly, the SCAD, adaptive LASSO, OSE, and LSA methods possess the so-called oracle property. This implies consistency in variable selection, namely, $\widehat{M}_k$ converges to the true M_k with probability tending to one when the sample size $n_k \rightarrow \infty$; and estimation

consistency; i.e., $\hat{\xi}_k$ converges in probability at rate $n_k^{-1/2}$ to the true value ξ_k^0 , say, and is asymptotically normal with the same mean and covariance matrix as the OLS estimator that would have been obtained had M_k been known in advance. As remarked previously, to first order, $\hat{\beta}$ should have the same asymptotic distribution as the estimator obtained by postulating models of form (10) and estimating ξ_k via standard methods with no variable selection. However, intuition suggests that use of variable selection methods with the oracle property in the above algorithm could be advantageous in practice.

4. Improved Sampling Variance Estimation

As noted in Sections 1 and 2, the usual sandwich estimator for the covariance matrix of the sampling distribution of any estimator for β solving (6), which we denote the “theoretical sandwich” (TS) estimator, can lead to optimistic standard errors and confidence intervals in practice, as seen in the simulations in Section 5. Tsiatis *et al.* [6] attributed this in part to failure of the estimator to take account of uncertainty due to fitting the treatment-specific regression models, which, although negligible for large n , could be problematic in finite samples. We now study this issue and propose an alternative approach that evidently takes account of some of this variation, focusing for definiteness on $K = 2$, with $\pi = \text{pr}(Z = 1)$.

Suppose that we have models for $E(Y|X, Z = k)$ of the form $q_k^*(X, \xi_k^*)$ for $k = 0, 1$, and write

$q_{k\xi}^*(X, \xi_k^*) = \partial/\partial \xi_k^* q_k^*(X, \xi_k^*)$. Let $\varpi = (\pi, \xi_0^{*T}, \xi_1^{*T}, \theta^T)^T$ denote the vector of all parameters. Suppose that ξ_0^* and ξ_1^* are estimated by OLS. Then the estimator $\hat{\varpi}$, say, solves jointly in ϖ the system of unbiased estimating equations

$$\begin{aligned} & \sum_{i=1}^n m^{**}(Y_i, X_i, Z_i; \varpi) \\ &= \sum_{i=1}^n \left[Z_i - \pi(1 - Z_i) \{Y_i - q_0^*(X_i, \xi_0^*)\} q_{0\xi}^*(X_i, \xi_0^*) Z_i \{Y_i - q_1^*(X_i, \xi_1^*)\} q_{1\xi}^*(X_i, \xi_1^*) m^*(Y_i, X_i, Z_i, \theta, \xi_0^*, \xi_1^*) \right] = 0, \end{aligned} \quad (12)$$

where $m^*(Y_i, X_i, Z_i, \theta, \xi_0^*, \xi_1^*)$ is equal to the summand in (6) with $E(Y_i|X_i, Z_i = k)$ replaced by $q_k^*(X_i, \xi_k^*)$ for $k = 0, 1$ and some fixed $m(Y, Z, \theta)$; and where we highlight the dependence of $m^*(Y_i, X_i, Z_i, \theta, \xi_0^*, \xi_1^*)$ on ξ_0^*, ξ_1^* . Then, letting $\Delta_i^{**}(\varpi) = E\{-\partial/\partial \varpi m^{**}(Y_i, X_i, Z_i, \varpi)\}$, the usual TS estimator for the covariance matrix of the estimator for θ solving (12) is the lower

right $(r \times r)$ submatrix of $\hat{\Delta}^{**1} \hat{\Omega}^{**} (\hat{\Delta}^{**1})^T$, where $\hat{\Delta}^{**} = n^{-1} \sum_{i=1}^n \Delta_i^{**}(\hat{\omega})$, $\hat{\Omega}^{**} = n^{-1} \sum_{i=1}^n m^{**}(Y_i, X_i, Z_i, \hat{\omega})^{\otimes 2}$, and $u^{\otimes 2} = uu^T$. Estimated variances for the estimator for β are then obtained from the relevant diagonal elements of this matrix; e.g., as in (9). As shown in the Supplementary Materials, terms in the matrix $\Delta_i^{**}(\hat{\omega})$ arising from the dependence of $m^{**}(Y_i, X_i, Z_i, \theta, \xi_0^*, \xi_1^*)$ on ξ_0^*, ξ_1^* are equal to zero, as they have expectation zero. The result is that the sampling covariance matrices of $(\hat{\xi}_0^{*T}, \hat{\xi}_1^{*T})^T$ and $\hat{\theta}$ are determined separately, so that the usual TS estimator does not involve terms related to estimation of ξ_0^*, ξ_1^* .

We consider replacing $\Delta_i^{**}(\hat{\omega})$ by $\tilde{\Delta}_i^{**}(\hat{\omega}) = -\partial/\partial\omega m^{**}(Y_i, X_i, Z_i, \hat{\omega})$ in the foregoing expressions, so that the terms arising from the dependence of $m^{**}(Y_i, X_i, Z_i, \theta, \xi_0^*, \xi_1^*)$ on ξ_0^*, ξ_1^* are no longer eliminated, and refer to the resulting covariance estimator as the “empirical sandwich” (ES). Although this will be inconsequential as $n \rightarrow \infty$, it may serve to take account of some variation due to estimation of ξ_0^*, ξ_1^* in finite samples. Evidence supporting this contention is presented in Section 5 and the Supplementary Material.

In case $K = 2$ and (1), it may be shown that the estimated sampling variance for $\hat{\beta}_1$ for $\beta_1 = E(Y|Z = 1) - E(Y|Z = 0)$ in (8) derived from this ES approach is given by

$$\begin{aligned} & \sum_{i=1}^n \left[\left\{ n_1^{-1} Z_i - n_0^{-1} (1 - Z_i) \right\} Y_i - n^{-1} \hat{\beta}_1 \right. \\ & \quad - (Z_i - \bar{Z}) \left\{ n_0^{-1} q_0^*(X_i, \hat{\xi}_0^*) + n_1^{-1} q_1^*(X_i, \hat{\xi}_1^*) \right\} \\ & \quad - (Z_i - \bar{Z}) \left\{ n_0^{-1} (\bar{Y}^{(0)} - \bar{q}_0^*) + n_1^{-1} (\bar{Y}^{(1)} - \bar{q}_1^*) \right\} \\ & \quad \left. + \left\{ n_1^{-1} \hat{D}_1 \hat{W}_1^{-1} \hat{\phi}_{1i} + n_0^{-1} \hat{D}_0 \hat{W}_0^{-1} \hat{\phi}_{0i} \right\} \right]^2, \end{aligned} \quad (13)$$

where $\hat{D}_k = n^{-1} \sum_{i=1}^n (Z_i - \bar{Z}) q_{k\xi}^{*T}(X_i, \hat{\xi}_k^*)$, $\hat{W}_0 = n^{-1} \sum_{i=1}^n (1 - Z_i) q_{0\xi}^*(X_i, \hat{\xi}_0^*) q_{0\xi}^{*T}(X_i, \hat{\xi}_0^*)$, $\hat{W}_1 = n^{-1} \sum_{i=1}^n Z_i q_{1\xi}^*(X_i, \hat{\xi}_1^*) q_{1\xi}^{*T}(X_i, \hat{\xi}_1^*)$, $\hat{\phi}_{1i} = Z_i \{Y_i - q_1^*(X_i, \hat{\xi}_1^*)\} q_{1\xi}^*(X_i, \hat{\xi}_1^*)$, and $\hat{\phi}_{0i} = (1 - Z_i) \{Y_i - q_0^*(X_i, \hat{\xi}_0^*)\} q_{0\xi}^*(X_i, \hat{\xi}_0^*)$. Comparing (13) to (9) shows that (13) involves an additional term whose form is clearly implicated in estimations of ξ_0^*, ξ_1^* .

To gain insight into why the ES approach yields more reliable standard error estimators in practice, we provide the following results, focusing on (1) with $K = 2$, so that $\beta_1 = E(Y|Z = 1) - E(Y|Z = 0)$ is of interest. Because the term in (9) and (13) having to do with estimation of π is identically zero if the models $q_k^*(X, \xi_k^*)$, $k=0, 1$ contain intercepts, we consider the case where π is held fixed at its known value rather than estimated, hence eliminating the first estimating equation in (12) to simplify the calculations.

Result 4.1

Consider (1) with $K = 2$. Assume correctly specified models $q_k^*(X, \xi_k^*)$ for $E(Y|X, Z = k)$, $k = 0, 1$, where ξ_k^0 is the true value of ξ_k^0 with ξ_k^0 estimated by the OLS estimator $\hat{\xi}_k^0$. Let $\sigma_\epsilon^2 = \text{Var}(Y|X, Z = k)$, $k=0, 1$, a constant. Let $\hat{\beta}_1$ be the estimator for $\beta_1 = E(Y|Z = 1) - E(Y|Z = 0)$, where β_1^0 is the true value. Then, under regularity conditions,

$$V^{TRUE} = \text{var} \left\{ n^{1/2} (\hat{\beta}_1 - \beta_1^0) \right\} = \text{var} \left\{ q_1^* (X_i, \xi_1^0) - q_0^* (X_i, \xi_0^0) \right\} + \frac{1}{\pi(1-\pi)} \sigma_\epsilon^2 + n^{-1} \sigma_\epsilon^2 \left\{ \frac{1-\pi}{\pi^2} p_1 + \frac{\pi}{(1-\pi)^2} p_0 \right\} + o(n^{-1}), \quad (14)$$

where p_0, p_1 are the numbers of elements in ξ_0^*, ξ_1^* , respectively.

A sketch of the argument is given in the Supplementary Material. Result 4.1 shows the form of the true variance of the estimator for β_1 to second order. The third term on the right hand side of (14), which will be negligible when n is large, may contribute to the true sampling variance in finite samples. Taken with the next result, Result 4.1 explains the optimism of the TS estimator and suggests why the ES estimator may perform better.

Result 4.2

Consider the situation of Result 4.1. Let V^{TS} and V^{ES} be the estimators for $\text{var} \left\{ n^{1/2} (\hat{\beta}_1 - \beta_1^0) \right\}$ obtained by the TS and ES approaches, respectively. Under regularity conditions,

$$E(V^{TS}) = V^{TRUE} - n^{-1} \left[\text{var} \left\{ q_1^* (X_i, \xi_1^0) - q_0^* (X_i, \xi_0^0) \right\} + \frac{1}{\pi(1-\pi)} \sigma_\epsilon^2 \right] + o(n^{-1}), \quad (15)$$

$$E(V^{ES}) = E(V^{TS}) + n^{-1} \sigma_\epsilon^2 \left\{ \frac{\pi}{(1-\pi)^2} p_0 + \frac{1-\pi}{\pi^2} p_1 \right\} + o(n^{-1}) \quad (16)$$

$$= V^{TRUE} - n^{-1} \left[\text{var} \left\{ q_1^* (X_i, \xi_1^0) - q_0^* (X_i, \xi_0^0) \right\} + \frac{1}{\pi(1-\pi)} \sigma_\epsilon^2 \right] + n^{-1} \sigma_\epsilon^2 \left\{ \frac{\pi}{(1-\pi)^2} p_0 + \frac{1-\pi}{\pi^2} p_1 \right\} + o(n^{-1}). \quad (17)$$

A sketch of the argument is in the Supplementary Material. Equation (15) shows that, although V^{TS} agrees in expectation with V^{TRUE} to first order, it will always be smaller to second order, so the TS approach may underestimate the true sampling variance in small samples and thus lead to optimistic Wald confidence intervals, as has been observed. Equation (16) shows that, while V^{TS} and V^{ES} agree in expectation to first order, as expected, the latter will always be larger to second order, so that the ES approach will result in larger estimated standard errors in small samples. The final expression (17) indicates, however, that V^{ES} could be larger or smaller than the true sampling variance in practice, and the extent to which V^{ES} comes close to the true sampling variance depends on the relative magnitudes of the second and third terms in (17).

These results do not address explicitly the potential effects of using variable selection methods to identify important variables. However, the form of (14) suggests that there may be a trade-off between gaining efficiency by decreasing $\text{var} \left\{ q_1^* (X_i, \xi_1^0) - q_0^* (X_i, \xi_0^0) \right\}$ and p_0, p_1 through a smaller model and by decreasing σ_ϵ^2 through a larger model. Result 4.2 does suggest that one approach to improved inference in finite samples in the special case of (1) with $K = 2$ would be to effect explicitly a second order correction to the TS estimator by adding an estimator of the second term on the right hand side of (15) to V^{TS} . As we show in the next section, using instead the ES approach directly to obtain standard errors yields good performance when combined with the variable selection algorithm and is easily generalized to other situations.

5. Simulation Studies

We have carried out extensive simulations to evaluate the performance of the proposed approach, focusing on $K = 2$. We report here on a subset of these studies involving continuous, binary, and longitudinal outcomes and various settings of factors including true model size, signal-to-noise ratio as reflected by the coefficient of determination R^2 , and sample size n ; additional results are in the Supplementary Material. We restricted attention to the typical situation in conventional clinical trials where $n > d$ and considered three classes of treatment effect estimators: the usual unadjusted estimator, traditional regression-based ANCOVA estimators, and semiparametric estimators using postulated treatment-specific regression models. In the latter two cases, we considered various approaches to model specification and variable selection, described next. This resulted in a large number of competing treatment effect estimators; for convenience, we group them into five categories.

Group I

The standard unadjusted estimator, denoted as “un.”

Group II

Usual ANCOVA estimators. We considered two types, ANCOVA1 (“an1”) and ANCOVA2 (“an2”) defined in [6], fitted in different ways; ANCOVA2 includes interactions of elements of X with Z . Given a definition of the total number of terms in X considered, e.g., all linear, quadratic, and two-way interactions, “-all” means include all terms, with no variable selection; “-linear” means include all d linear terms only, with no variable selection (same as “-all” if only linear terms considered); “-missub” means include a prespecified subset of terms, representing the situation where model and covariates are dictated in the protocol; and “-forw- α ” using classical forward selection with “ $\alpha = 0.05$ to enter” to select from among the considered terms. For linear or logistic models, forward selection was implemented using the leaps package, available at <http://cran.r-project.org/>, or the step function in R.

Group III

The semiparametric estimator with no variable selection used to fit assumed models $q_k(X, \xi_k)$, $k = 0, 1$, as in (10), with $c_k(X)$ involving s_k terms, including all s_k terms in model k , denoted “semi-all.”

Group IV

The semiparametric estimator (“semi”) with various forms of forward selection used to estimate M_k in assumed models $q_k(X, \xi_k)$ considering all s_k terms. “-forw- α ” means classical forward selection “ $\alpha = 0.05$ to enter;” “-forw-aic” means forward selection tuned by using the Akaike Information Criterion (AIC); “-forw-bic” means forward selection tuned by BIC, all implemented using the leaps package (linear models) or step function (logistic models) in R; and “-forw-fsr” means forward selection with Fast FSR tuning as in [13] with target FSR = 0.05, implemented using R code available at <http://www4.stat.ncsu.edu/~boos/var.select/>.

Group V

The semiparametric estimator (“semi”) with various shrinkage methods used for variable selection to estimate M_k in postulated models $q_k(X, \xi_k)$, considering all s_k terms. “-lasso” means LASSO selection with BIC used to choose the tuning parameters; “-adallasso” means adaptive LASSO selection with BIC tuning; “-ose- $L_{0.01}$ ” means one-step sparse selection with penalty $|\xi|^{0.01}$, tuned by BIC; “-ose-scad” means one-step selection with SCAD penalty, tuned by BIC; and “-scad” means SCAD selection with BIC tuning. The first

four methods were implemented using the lars (linear models) or glmnet (logistic models) R packages; the last was implemented using the ncvgreg package in R, all available at <http://cran.r-project.org/>.

As a benchmark, we considered the semiparametric estimator with $E(Y|X, Z = k)$, $k = 0, 1$, correctly specified, including only important terms in the fit, denoted “semi-oracle.”

5.1. Continuous outcome

Our first scenario is a simple setting with $\text{pr}(Z = 0) = \text{pr}(Z = 1) = 0.5$ and $d = 20$, $X_1, X_3, X_8, X_9, \dots, X_{20}$ iid $N(0, 1)$; X_4, X_6 Bernoulli with $\text{pr}(X_4 = 1) = 0.3$ and $\text{pr}(X_6 = 1) = 0.5$; and $X_2 = 0.20X_1 + 0.98U_1$, $X_5 = 0.10X_1 + 0.20X_3 + 0.97U_2$, $X_7 = 0.10X_3 + 0.99U_3$, with U_1, U_2, U_3 iid $N(0, 1)$. Outcomes were generated as $Y_i = (\zeta_{00} + \zeta_{01}^T X_i) Z_i + (\zeta_{10} + \zeta_{11}^T X_i) (1 - Z_i) + \epsilon_i$, with ϵ_i iid $N(0, \sigma^2)$. Here, $\beta_1 = E(Y|Z = 1) - E(Y|Z = 0) = \zeta_{10} - \zeta_{00} + (\zeta_{11} - \zeta_{01})^T E(X)$. We chose $\zeta_{00}, \zeta_{01}, \zeta_{10}, \zeta_{11}$ to yield mild, moderate, and strong relationships, reflected by the magnitudes of R^2 in each treatment group. We report here on the moderate case with $n = 120$ and $\sigma = 1.0$; results for other scenarios are given in the Supplementary Material. Here, $\zeta_{00} = -0.3$, $\zeta_{01} = (0, 0, 0, 0.5, 1, 1, 0.1, 0, \dots, 0)^T$, $\zeta_{10} = 0.6$, and $\zeta_{11} = (0.2, 0.5, 0.5, 0.5, 0.5, 0.4, 0, \dots, 0)^T$ so that $\beta_1^0 = 0.6$ and $R^2 = 0.59, 0.55$ for groups 0, 1, respectively. For the ANCOVA methods, we considered only linear terms in elements of X ; “an1-missub” and “an2-missub” include only X_2, X_7 . For the semiparametric methods, we took the assumed models to be of form $q_k(X, \xi_k) = \xi_{k0} + \xi_{ks}^T X$, so that $s_k = d$, $k = 0, 1$, and the models contain the truth. 5000 Monte Carlo data sets were used.

Results are in Table 1. All estimators show negligible bias, as expected. Both traditional ANCOVA and semiparametric approaches show efficiency gains over the unadjusted estimator, except in the case of ANCOVA with a prespecified model, which fails to include important covariates and hence provides ineffective adjustment. Including all 20 covariates in either the ANCOVA or semiparametric approaches yields gains; however, these can be improved upon by 10% or more by using some form of variable selection and the semiparametric methods. More dramatic gains are possible in settings with stronger covariate relationships; see the Supplementary Material. Overall, the qualitative impression, duplicated over all other simulations we have carried out with different n and R^2 , is that using adaptive LASSO and one-step sparse selection in Group V and Fast FSR in Group IV yield semiparametric estimators that achieve consistently large efficiency gains. No estimator attains the performance of the benchmark, demonstrating that the need for variable selection exacts a price for small and moderate n ; however, the modern methods consistently come closest to this ideal.

Comparing the average of estimated standard errors obtained from V^{TS} and V^{ES} to the Monte Carlo standard deviation shows that the former consistently underestimate true sampling variation, leading to coverage probabilities of corresponding Wald-type 95% confidence intervals that do not achieve the nominal level (with 5000 replicates, standard error for Monte Carlo coverage is 0.003). When used with the shrinkage or Fast FSR methods, standard errors based on V^{ES} consistently yield reliable estimation of uncertainty and nominal coverage. We observed this behavior for small and moderate n across all simulations we carried out.

Table 2 shows treatment-specific variable selection statistics across the 5000 simulated data sets. The true model was rarely identified perfectly by any method owing to the mild effects of some variables, making them hard to detect; in scenarios with stronger relationships and larger n in the Supplementary Material, much higher proportions of correct identification are possible. Taken with Table 1, these results show that the methods achieving the best

efficiency gains were also best at identifying components of X with true zero coefficients, even at the expense of eliminating one or two important variables. Evidently, achieving parsimonious models containing few irrelevant terms is implicated in improved efficiency in practice.

To gauge performance in the situation of a typical large clinical trial, we considered again $\text{pr}(Z = 0) = \text{pr}(Z = 1) = 0.5$ and $d = 20$, with $n = 1200$. Interest again focused on $\beta_1 = E(Y|Z = 1) - E(Y|Z = 0)$. Here, $X_1, \dots, X_{15}$ were iid $N(0, 1)$; X_{16}, X_{17}, X_{18} were each Bernoulli with success probabilities 0.3, 0.5, 0.7, respectively; and X_{19} (X_{20}) was binomial with 5 (3) trials and success probability 0.4 (0.8). Outcomes were generated via $Y = 0.1 + 0.4X_2 + 0.5X_3 + 0.7X_{16} + 0.4X_{19} + 0.8X_3^2 + 0.9X_3X_{16} + \epsilon$ and $Y = -0.1 + X_2 + X_4 + X_{17} + X_{20} + X_4^2 + X_2X_{17} + X_2X_{20} + \epsilon$ for treatments 0 and 1, respectively, where $\epsilon \sim N(0, 1)$, which yields $\beta_1^0 = 1.89$ and $R^2 = 0.39, 0.55$ for treatments 0, 1. For the ANCOVA methods, we considered all linear and two-way interaction terms in elements of X along with quadratic terms in all continuous covariates, for a total of 225 potential terms, yielding a high-dimensional variable selection problem. Here, “an1-missub” and “an2-missub” included only linear terms in X_2, X_7 . For the semiparametric methods, we took the assumed models to be of the form

$q_k(X, \xi_k) = \xi_{k0} + \xi_{ks}^T c_k(X)$ with $c_k(X)$ containing all $s_k = 225$ components above, $k = 0, 1$. Thus, variable selection methods must identify from a large number of potential terms relatively sparse models.

The results in Table 3 show that, again, all estimators exhibit negligible bias, with the semiparametric methods using Fast FSR or shrinkage type selection achieving larger efficiency gains over the unadjusted estimator than classical approaches (ANCOVA, classical “ α -to-enter” forward selection), approaching the benchmark value. With the large sample size, the difference between standard errors obtained using V^{TS} and V^{ES} is negligible, as expected, and coverage probabilities for the semiparametric methods with Fast FSR or shrinkage type selection achieve the nominal level. As in the first scenario, Fast FSR, adaptive LASSO, and one-step sparse selection perform well.

5.2. Binary outcome

We considered binary outcome and (2) with $K = 2$, so that, taking $k = 0, 1$, interest focuses on β_1 in $\text{logit}\{E(Y|Z)\} = \beta_0 + \beta_1 Z$; i.e., $\beta_1 = \text{logit}\{E(Y|Z = 1)\} - \text{logit}\{E(Y|Z = 0)\}$.

With $\text{pr}(Z = 0) = \text{pr}(Z = 1) = 0.5$, we took $d = 8$, X_1, X_3, X_8 iid $N(0, 1)$; X_4, X_6 Bernoulli with $\text{pr}(X_4 = 1) = 0.3$, $\text{pr}(X_6 = 1) = 0.5$; and $X_2 = 0.20X_1 + 0.98U_1$, $X_5 = 0.10X_1 + 0.20X_3 + 0.97U_2$, $X_7 = 0.10X_3 + 0.99U_3$, with U_1, U_2, U_3 iid $N(0, 1)$. Binary outcomes Y were generated as Bernoulli with success probabilities according to $\text{logit}\{E(Y|X, Z)\} =$

$(\zeta_{00} + \zeta_{01}^T X)Z + (\zeta_{10} + \zeta_{11}^T X)(1 - Z)$. As in the continuous case, we specified different $(\zeta_{00}, \zeta_{01}, \zeta_{10}, \zeta_{11})$ to yield varying strength of relationships, and report here on a case with relatively small $n = 100$ and $R^2 = 0.33, 0.37$ for treatments 0, 1, achieved by taking $\zeta_{00} = -0.80$, $\zeta_{01} = (1.2, 1, 0, 0, 0, 0, 0, 0)^T$, $\zeta_{10} = 0.38$, and $\zeta_{11} = (0.5, 1.3, 0.5, 1.5, 0, 0, 0, 0)^T$.

For this set-up, $\beta_1^0 = -0.49$. Results are in Table 4; other cases are presented in the Supplementary Material.

The usual Group II ANCOVA approaches in this setting would be to fit by IRWLS a conditional (on X) model of the form $\text{logit}\{E(Y|X, Z)\} = \beta_0^* + X^T \phi^* + \beta_1^* Z$, say; in Table 4, we show results for estimation of β_1^* for various versions of this, denoted “an1,” where we considered only linear terms in X , in the “logistic models” section of the table. It is important to recognize that, in contrast to the situation with linear models, β_1 in the model of interest for $E(Y|Z)$ is different from β_1^* in this model. Thus, as discussed by [6, 19] and

others, such ANCOVA approaches are expected to yield inconsistent estimators for the true treatment effect. This is clearly the case for various versions in Table 4, which consequently exhibit abysmal efficiency relative to the unadjusted estimator.

As noted in Section 2, the semiparametric theory guarantees increased efficiency of estimation of β if the postulated models $q_k(X, \xi_k)$, $k = 0, 1$, are taken to be linear in ξ_k and fitted by OLS. We thus considered two versions of the semiparametric estimator: those where we took $\text{logit}\{q_k(X, \xi_k)\} = \{q_k(X, \xi_k)\} = \xi_{k0} + \xi_{ks}^T X$, $k=0, 1$, the correct form, with fitting by IRWLS, results for which are listed in the “logistic models” section of Table 4; and those using linear models $q_k(X, \xi_k) = \xi_{k0} + \xi_{ks}^T X$, $k=0, 1$, with OLS, shown in the “linear models” section, so that in either case $s_k = d$. We take as the ideal benchmark the estimator based on the correct logistic models (“semi-oracle”); for the linear models, we include as a “benchmark” the case where the true important covariates are included in the linear models (“semi-correctx”).

From Table 4, the semiparametric estimators using either logistic or linear models all show negligible bias and substantial efficiency gains over the unadjusted. Those using either type of model including all covariates or using classical forward selection with “ α -to-enter” do not achieve the same gains as the others. In general, logistic models result in slightly better performance; however, coverage probabilities are still short of the nominal level using either standard error estimator. The relative efficiency loss incurred using linear models is very small, and confidence intervals based on V^{ES} achieve nominal coverage. As in the continuous case, the adaptive LASSO and one-step sparse selection methods perform well. Thus, given the increased computational cost of variable selection for logistic models relative to the linear case, and as suggested by the semiparametric theory, using linear models in this context with these variable selection methods may be attractive in practice.

5.3. Longitudinal outcome

We considered the scenario in [7], (3) with $K = 2$, so, using $k = 0, 1$, interest is in β_1 in the linear mixed model $Y_{ij} = \tau_0 + (\beta_0 + \beta_1 Z_i)t_{ij} + b_{0i} + b_{1i}t_{ij} + e_{ij}$, $i = 1, \dots, n$, $j = 1, \dots, m_i$, $\beta = (\beta_0, \beta_1)^T$. We took $n = 200$; the numbers of observations per subject $m_i = 9, 10, 11$ with equal probability; $t_{ij} = 2(j - 1)$, $j = 1, \dots, m_i$; Z Bernoulli with $\text{pr}(Z = 0) = \text{pr}(Z = 1) = 0.5$; and X_1, X_3 iid $N(0, 1)$, $X_2 = 0.20X_1 + 0.96U_2$, $U_2 \sim N(0, 1)$, so $d = 3$. Longitudinal outcomes were generated according to $Y_{ij} = \eta_{i0} + \eta_{1i}t_{ij} + e_{ij}$, where $\eta_{0i} = 0.5 + 0.2X_{1i} + 0.5X_{2i} + b_{0i}$; for i in treatment group k , $\eta_{1i} = \zeta_{0k} + \zeta_{1k}X_{1i}^2 + \zeta_{2k}X_{2i} + \zeta_{3k}X_{3i} + b_{1i}$, with $(b_{0i}, b_{1i}) \sim N(0, D)$ for (2×2) covariance matrix D with $D_{11} = 1$, $D_{12} = 0.2$, $D_{22} = 0.4$, so that $\text{corr}(b_{0i}, b_{1i}) = 0.32$; and e_{ij} were iid $N(0, \sigma_e^2 = 16)$. We used different values of $\zeta_k = (\zeta_{0k}, \zeta_{1k}, \zeta_{2k}, \zeta_{3k})^T$ to achieve different strengths of relationships; we report here on the case where $\zeta_0 = (1.0, 0.2, 0.2, 0)^T$ and $\zeta_1 = (1.0, 0.2, 0, 0.2)^T$, yielding $R^2 = 0.24$ in each group and true $\beta_1^0 = 0.3 = 0.3$. Results for other cases are given in the Supplementary Material.

For each of 5000 Monte Carlo data sets, we obtained by the unadjusted estimate by fitting (3) directly. Analogous to ANCOVA in this setting, we fitted the conditional (on X) linear mixed model $Y_{ij} = \tau_0^* + \phi_0^{*T} X_i + (\beta_0^* + \phi_1^{*T} X_i + \beta_1^* Z_i) t_{ij} + b_{0i} + b_{1i}t_{ij} + e_{ij}$, with linear covariate effects only, as might be dictated in a trial protocol, denoted by “mixed-linear” in Table 5. The unadjusted estimator involves solving estimating equations in

$\theta = (\beta^T, \tau_0, \text{vech}(D)^T, \sigma_e^2)^T$; those for τ_0, β^T are of form (5), while those for the remaining elements are uncorrelated with these but quadratic in the outcomes. Following [7], for the semiparametric estimators, we fixed $\text{vech}(D), \sigma_e^2$ in $m(Y, Z, \theta)$ at the estimates from the

unadjusted analysis and estimated $(\tau_0, \beta^T)^T$ only. As did these authors, we postulated models $q_k(X, \xi_k)$, $k = 0, 1$, for scalar $E(Y_{ij}|X_i, Z_i = k)$ containing all linear, quadratic, and two-way interactions in elements of X and interactions of these with time t_{ij} , so that $s_k = 18$. The benchmark uses the true $E(Y_{ij}|X_i, Z_i = k)$ implied by the model generating the data.

From Table 5, all estimators show negligible bias. The “usual” ANCOVA-like analysis yields only a modest efficiency gain over the unadjusted. The semiparametric approach including all s_k terms offers no improvement. In contrast, using some form of variable selection performs well, with the Fast FSR and shrinkage methods best. Use of V^{ES} to obtain standard errors leads to nominal coverage in these cases.

6. Application

We demonstrate the methods by application to data on CD4 count (cells/mm³, Y) at 20 ± 5 weeks post-baseline from 2139 subjects from AIDS Clinical Trials Group (ACTG) protocol 175 [20]. As in [6], we consider the group zidovudine (ZDV) monotherapy ($Z = 0$, $n_0 = 532$) and group combining the three other treatment arms ZDV+didanosine (ddI), ZDV+zalcitabine, and ddI monotherapy ($Z = 1$, $n_1 = 1607$); interest focuses on $\beta_1 = E(Y|Z = 1) - E(Y|Z = 0)$. The covariates X included continuous variables baseline CD4 count (cells/mm³), baseline CD8 count (cells/mm³), age (years), weight (kg) and Karnofsky score (0–100), and binary indicators for hemophilia, homosexual activity, history of drug use, race (1=non-white), gender (0=female), antiretroviral history (0=naive), and symptomatic status (0=asymptomatic). For each group, we took $q_k(X, \xi_k)$ to contain all linear terms, quadratic terms in the five continuous variables, and most two-way interactions, for a total of $s_k = 83$ potential terms. Results are shown in Table 6 and agree with those in Table I of [6] for “un,” “an1-linear,” and “semi-forw- α .” The ANCOVA “missub” methods included linear terms in age and Karnofsky score only. Evidently, use of the semiparametric methods with some form of tuned variable selection yields the most precise inferences.

7. Discussion

We have investigated performance of the semiparametric methods for covariate adjusted analysis of clinical trials proposed by [6, 7] when classical and modern variable selection methods are used to develop treatment-specific outcome regression models that are used to effect the adjustment. In the form of the methods considered here, these models may be developed independently of estimation of the treatment effect; thus, when integrated with variable selection procedures, they offer the opportunity for a fully objective and transparent approach to covariate adjustment that may be prespecified in a trial protocol. The semiparametric methods integrated with modern variable selection techniques outperform traditional ANCOVA using classical model selection on the basis of gains in efficiency relative to an unadjusted analysis. Given availability of software to implement the variable selection methods, which we used in our numerical studies, the approach is straightforward to adopt and duplicate in practice. Although we focused on two treatment arms, the approach and qualitative conclusions extend readily to more than two treatments.

Our results support the intuition that variable selection procedures that excel at identifying true important variables and regression relationships, including shrinkage methods possessing selection consistency, lead to the most precise adjusted estimators of treatment effect. This is particularly true when the number of covariates is large and the underlying true treatment-specific regression relationships are sparse (i.e., involve just a few variables). Across our studies, the Fast FSR, one-step, and adaptive LASSO methods performed consistently well in this setting. However, when there are only weak signals associated with several covariates, the LASSO might be preferred; while it does not possess the oracle

property, it may outperform these other approaches in this setting. Considering performance and simplicity of implementation, if a single method is to be recommended, we suggest the adaptive LASSO based on overall resulting efficiency of treatment effect estimation and ease of implementation in available software. In situations where few covariates are available, modern or classical methods are likely to yield similar results, and hence the choice may be based on ease of implementation.

We have also proposed an approach to obtaining analytical approximate standard errors for the resulting treatment effect estimators that yields confidence intervals for the true treatment effect that achieve the nominal level when used with linear treatment-specific models and selection-consistent methods such as the adaptive LASSO. This approach offers improved performance over the usual sandwich method. A natural alternative would be to employ a standard bootstrap technique; however, as warned by [6], traditional bootstrapping may not be valid when the use of variable selection methods to deduce the treatment-specific regression is involved. We have confirmed this conjecture in empirical studies and have developed an alternative, valid resampling technique that yields performance comparable to that of the empirical sandwich methods proposed here; this work will be reported elsewhere.

For binary outcome, the treatment-specific models may be developed as nonlinear binary regression models, e.g., logistic, or as linear models; software implementing adaptive LASSO selection is available for both. The latter are guaranteed in large samples by semiparametric theory to result in efficiency gains if fitted by OLS, but our studies indicate that logistic models, which may contain the true regression relationships, can lead to additional improvements in efficiency, as intuition would suggest, at the expense of increased computational cost when the number of potential model terms is large. However, our studies also show that estimation of uncertainty and confidence intervals may be more reliable in small/moderate samples if linear models are used. We thus recommend using linear models in this case and when the number of potential terms is large; in larger samples, logistic models may be preferred. As the models are of interest only for their role in increasing efficiency of estimation of the treatment effect, there is no conceptual issue with using linear models for adjustment.

A common concern in practice regarding the use of covariate adjustment of any type is whether or not the available baseline covariate information is sufficiently prognostic to make use of an adjusted estimator over the usual unadjusted method worthwhile. The semiparametric theory guarantees that, if one uses linear treatment-specific models fitted by OLS, regardless of whether or not they are correct, the semiparametric estimator will be at least as efficient as the unadjusted in large samples. Accordingly, in very large trials where the uncertainty due to model-fitting and variable selection is likely to be negligible, a reasonable strategy might be to favor routinely the semiparametric estimator with variable selection as discussed here. More generally, a principled diagnostic approach to resolving this issue is not available. Our extensive simulations suggest that, not surprisingly, the efficiency gains achieved by the semiparametric estimator are related to the magnitudes of the proportion of variation in the outcome explained by the covariates as reflected in the treatment-specific coefficients of determinations R^2 , with values of $R^2 > 0.25$ being associated with nonnegligible gains. A heuristic tactic to be used with the approach here might thus be to inspect the R^2 values associated with the selected models, using this rule of thumb. In some subject-matter contexts, prognostic covariates may be expected; see [10] for a case where adjustment would be advocated routinely.

Semiparametric treatment effect estimators have been derived in the case of censored survival data [6, 11]; owing to the additional complexity introduced by the censoring, we did

not consider this situation here. The use of variable selection approaches in this context is an open problem.

We have focused on the semiparametric approach as proposed in [6, 7], in which the outcome-covariate relationship is explicitly modeled separately by treatment, facilitating implementation by distinct groups of analysts if full transparency is required. Tian *et al.* [21] proposed an approach to variable selection and sampling variance estimation based on shrinkage methods and cross-validation within the class of semiparametric estimators where a model for $E(Y|X, Z)$ is taken to be linear in X and Z and is fitted simultaneously with the treatment effect. In a related context under similar conditions, [19] proposed a variable selection procedure focused on optimizing a measure of the efficiency to be gained by covariate adjustment.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by NIH grants P01 CA142538, R01 CA085848, and R37 AI031789 and NSF grant DMS-0645293.

References

1. Koch GG, Tangen CM, Jung JW, Amara IA. Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and non-parametric strategies for addressing them. *Statistics in Medicine*. 1998; 17:1863–1892. DOI: 10.1002/(SICI)1097-0258(19980815/30)17:15/16<1863::AID-SIM989>3.0.CO;2-M. [PubMed: 9749453]
2. Assmann EE, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet*. 2000; 355:1064–1069. DOI:10.1016/S0140-6736(00)02039-0. [PubMed: 10744093]
3. Pocock SJ, Assmann SE, Enos LL, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: Current practice and problems. *Statistics in Medicine*. 2002; 21:2917–2930. DOI: 10.1002/sim.1296. [PubMed: 12325108]
4. Lesaffre E, Senn S. A note on non-parametric ANCOVA for covariate adjustment in randomized clinical trials. *Statistics in Medicine*. 2003; 22:3583–3596. DOI: 10.1002/sim.1583. [PubMed: 14652862]
5. Grouin JM, Day S, Lewis J. Adjustment for baseline covariates: an introductory note. *Statistics in Medicine*. 2004; 23:697–699. DOI: 10.1002/sim.1646. [PubMed: 14981669]
6. Tsiatis AA, Davidian M, Zhang M, Lu X. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine*. 2008; 27:4658–4677. DOI: 10.1002/sim.3113. [PubMed: 17960577]
7. Zhang M, Tsiatis AA, Davidian M. Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics*. 2008; 64:707–715. DOI: 10.1111/j.1541-0420.2007.00976.x. [PubMed: 18190618]
8. Tsiatis, AA. *Semiparametric Theory and Missing Data*. Springer; New York: 2006.
9. Lu X, Tsiatis AA. Improving efficiency of the log-rank test using auxiliary covariates. *Biometrika*. 2008; 95:676–694. DOI: 10.1093/biomet/asn003.
10. Gilbert PB, Sato M, Sun X, Mehortha DV. Efficient and robust method for comparing the immunogenicity of candidate vaccines in randomized clinical trials. *Vaccine*. 2009; 27:396–401. DOI: 10.1016/j.vaccine.2008.10.083. [PubMed: 19022314]
11. Zhang M, Gilbert PB. Increasing the efficiency of prevention trials by incorporating baseline covariates. *Statistical Communications in Infectious Diseases*. 2010; 2:1–28. DOI: 10.2202/1948-4690.1002. [PubMed: 21152074]

12. Carroll, R.J.; Ruppert, D.; Stefanski, L.A.; Crainiceanu, C.M. *Measurement Error in Nonlinear Models: A Modern Perspective*. Second Edition. Chapman and Hall/CRC; Boca Raton: 2006.
13. Boos DD, Stefanski LA, Wu Y. Fast FSR variable selection with applications to clinical trials. *Biometrics*. 2008; 65:692–700. DOI: 10.1111/j.1541-0420.2008.01127.x. [PubMed: 18945266]
14. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*. 1996; 58:267–288.
15. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle property. *Journal of the American Statistical Association*. 2001; 96:1348–1360. DOI: 10.1198/016214501753382273.
16. Zou H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*. 2006; 101:1418–1429. DOI: 10.1198/016214506000000735.
17. Zou H, Li R. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*. 2008; 36:1509–1533. DOI: 10.1214/009053607000000802. [PubMed: 19823597]
18. Wang H, Leng C. Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association*. 2007; 102:1039–1048. DOI: 10.1198/016214507000000509.
19. Moore KL, Neugebauer R, Valappil T, van der Laan MJ. Robust extraction of covariate information to improve estimation efficiency in randomized clinical trials. *Statistics in Medicine*. 2011; 30:2389–2408. DOI: 10.1002/sim.4301. [PubMed: 21751231]
20. Hammer SM, Katzenstein DA, Hughes MD, Gundaker H, Schooley RT, Haubrich RH, Henry WK, Lederman MM, Phair JP, Niu M, Hirsch MS, Merigan TC, for the AIDS Clinical Trials Group Study 175 Study Team. A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*. 1996; 335:1081–1089. [PubMed: 8813038]
21. Tian L, Cai T, Zhao L, Wei LJ. On the covariate-adjusted estimation for an overall treatment difference with data from a randomized comparative clinical trial. *Biostatistics*. 2012; 13:256–273. DOI: 10.1093/biostatistics/kxr050. [PubMed: 22294672]

Table 1

Simulation results for estimation of $\beta_1 = E(Y|Z = 1) - E(Y|Z = 0)$, first continuous outcome scenario, with $n = 120$ and true $\beta_1^0 = 0.6$, 5000 Monte Carlo replications. Bias, mean of Monte Carlo bias; MCSD, Monte Carlo standard deviation; SE^{TS} , average of standard error estimates using the “theoretical sandwich” formula; CP^{TS} , Monte Carlo coverage of corresponding 95% Wald confidence intervals; SE^{ES} , average of standard error estimates using the “empirical sandwich” formula; CP^{ES} , Monte Carlo coverage of corresponding 95% Wald confidence intervals; RE, Monte Carlo mean square error for the unadjusted methods divided by that for the indicated method.

Group	Method	Bias	RE	SE^{TS}	SE^{ES}	MCSD	CP^{TS}	CP^{ES}
I	un	−0.005	1.000	0.270	–	0.269	0.950	–
II	an1-all	−0.004	1.443	0.219	–	0.224	0.941	–
	an1-missub	−0.006	1.012	0.266	–	0.268	0.947	–
	an1-forw- α	−0.005	1.505	0.201	–	0.219	0.924	–
	an2-all	−0.003	1.422	0.205	–	0.226	0.920	–
	an2-missub	−0.005	1.013	0.261	–	0.267	0.942	–
	an2-forw- α	−0.005	1.525	0.179	–	0.218	0.895	–
III	semi-all	−0.003	1.414	0.187	0.202	0.226	0.894	0.917
IV	semi-forw- α	−0.005	1.554	0.192	0.211	0.216	0.916	0.935
	semi-forw-aic	−0.005	1.515	0.190	0.206	0.219	0.908	0.929
	semi-forw-bic	−0.005	1.574	0.194	0.213	0.215	0.919	0.938
	semi-ffsr	−0.004	1.576	0.200	0.221	0.214	0.926	0.951
V	semi-scad	−0.005	1.568	0.194	0.213	0.215	0.919	0.940
	semi-lasso	−0.005	1.553	0.202	0.223	0.216	0.931	0.952
	semi-adalasso	−0.005	1.603	0.197	0.216	0.213	0.929	0.946
	semi-ose- $L_{0.01}$	−0.005	1.603	0.197	0.216	0.213	0.929	0.946
	semi-ose-scad	−0.005	1.572	0.195	0.214	0.215	0.921	0.946
benchmark	semi-oracle	−0.005	1.697	0.195	0.215	0.207	0.930	0.953

Table 2

Variable selection results for continuous outcome with $n = 120$, and moderate R^2 , 5000 Monte Carlo replications. CZ_0 , CZ_1 , numbers of zero coefficients estimated to be zero in the selected models for treatments 0,1; ICZ_0 , ICZ_1 , numbers of nonzero coefficients incorrectly estimated to be zero in the selected models for treatments 0,1; $CFIT_0$, $CFIT_1$, proportions of the 5000 data sets for which the true models were selected for treatments 0,1.

Group	Method	CZ_0	ICZ_0	$CFIT_0$	CZ_1	ICZ_1	$CFIT_1$
IV	semi-forw- α	4.21	1.34	0.02	12.44	1.90	0.01
	semi-forw-aic	12.27	1.07	0.01	10.71	1.36	0.01
	semi-forw-bic	14.90	1.50	0.02	12.97	2.20	0.01
	semi-ffsr	15.74	1.92	0.01	13.68	3.03	0.00
V	semi-scad	14.89	1.54	0.02	12.93	2.22	0.01
	semi-lasso	14.64	1.55	0.01	12.36	1.94	0.01
	semi-adalasso	14.96	1.60	0.01	12.68	2.08	0.01
	semi-ose- $L_{0.01}$	14.96	1.60	0.01	12.68	2.08	0.01
	semi-ose-scad	14.88	1.59	0.01	12.79	2.22	0.01
benchmark	semi-oracle	16.00	0.00	1.00	14.00	0.00	1.00

Table 3

Simulation results for estimation of $\beta_1 = E(Y|Z = 1) - E(Y|Z = 0)$, second continuous outcome scenario, with $n = 1200$ and true $\beta_1^0 = 1.89$, 1000 Monte Carlo replications. Entries are as in Table 1.

Group	Method	Bias	RE	SE^{TS}	SE^{ES}	MCS D	CP^{TS}	CP^{ES}
I	un	0.000	1.000	0.266	–	0.274	0.947	–
II	an1-all	–0.002	1.196	0.245	–	0.251	0.945	–
	an1-missub	–0.006	1.332	0.235	–	0.237	0.944	–
	an1-forw- α	–0.008	1.337	0.217	–	0.237	0.930	–
	an2-all	–0.002	1.177	0.213	–	0.253	0.908	–
	an2-missub	–0.006	1.335	0.212	–	0.237	0.931	–
	an2-forw- α	–0.005	1.373	0.174	–	0.234	0.862	–
III	semi-all	–0.002	1.175	0.213	0.229	0.253	0.912	0.930
IV	semi-forw- α	–0.004	1.381	0.217	0.218	0.233	0.936	0.938
	semi-forw-aic	–0.005	1.317	0.214	0.217	0.239	0.919	0.922
	semi-forw-bic	–0.003	1.451	0.220	0.221	0.228	0.940	0.941
	semi-ffsr	–0.002	1.463	0.222	0.222	0.227	0.947	0.946
V	semi-scad	–0.003	1.433	0.222	0.223	0.229	0.945	0.944
	semi-lasso	–0.004	1.435	0.223	0.224	0.229	0.951	0.952
	semi-adalasso	–0.004	1.445	0.223	0.223	0.228	0.951	0.951
	semi-ose- $L_{0.01}$	–0.004	1.445	0.223	0.223	0.228	0.951	0.951
	semi-ose-scad	–0.005	1.427	0.223	0.223	0.228	0.946	0.946
benchmark	semi-oracle	–0.004	1.473	0.221	0.222	0.226	0.949	0.949

Table 4

Simulation results for for estimation of $\beta_1 = \text{logit}\{E(Y|Z=1)\} - \text{logit}\{E(Y|Z=0)\}$, binary outcome scenario, with $n = 100$ and true $\beta_1^0 = -0.490$, 5000 Monte Carlo replications. Entries are as in Table 1.

Group	Method	Bias	RE	SE^{TS}	SE^{ES}	MCSD	CP^{TS}	CP^{ES}
Logistic models								
I	un	-0.014	1.000	0.410	-	0.409	0.950	-
II	an1-all	-0.341	0.337	0.565	-	0.616	0.919	-
	an1-missub	-0.156	0.650	0.474	-	0.482	0.950	-
	an1-forw- α	-0.283	0.416	0.524	-	0.567	0.925	-
III	semi-all	-0.009	1.267	0.325	0.334	0.363	0.918	0.922
IV	semi-forw- α	-0.012	1.169	0.363	0.469	0.378	0.942	0.958
	semi-forw-aic	-0.011	1.288	0.330	0.344	0.360	0.926	0.933
	semi-forw-bic	-0.013	1.295	0.337	0.354	0.359	0.931	0.940
V	semi-scad	-0.013	1.293	0.335	0.352	0.359	0.930	0.939
	semi-lasso	-0.015	1.266	0.344	0.360	0.363	0.934	0.947
	semi-adalasso	-0.015	1.289	0.335	0.350	0.394	0.928	0.940
	semi-ose- $L_{0.01}$	-0.013	1.290	0.335	0.351	0.360	0.927	0.938
	semi-ose-scad	-0.012	1.288	0.336	0.352	0.360	0.928	0.939
benchmark	semi-oracle	-0.012	1.379	0.336	0.354	0.348	0.940	0.951
Linear models								
III	semi-all	-0.018	1.248	0.334	0.348	0.366	0.927	0.936
IV	semi-forw- α	-0.018	1.256	0.344	0.360	0.364	0.936	0.948
	semi-forw-aic	-0.018	1.274	0.337	0.352	0.362	0.932	0.943
	semi-forw-bic	-0.017	1.279	0.342	0.358	0.361	0.939	0.948
	semi-ffsr	-0.017	1.272	0.345	0.362	0.362	0.939	0.949
V	semi-scad	-0.017	1.277	0.342	0.359	0.361	0.939	0.948
	semi-lasso	-0.017	1.264	0.349	0.366	0.363	0.940	0.953
	semi-adalasso	-0.017	1.282	0.344	0.360	0.361	0.940	0.949
	semi-ose- $L_{0.01}$	-0.017	1.282	0.344	0.360	0.361	0.940	0.949
	semi-ose-scad	-0.017	1.281	0.342	0.359	0.361	0.939	0.949
benchmark	semi-correctx	-0.017	1.351	0.341	0.357	0.351	0.944	0.953

Table 5

Simulation results for estimation of β_1 in the linear mixed model, longitudinal outcome scenario, with $n = 200$ and true $\beta_1^0 = 0.3$, 5000 Monte Carlo replications. Entries are as in Table 1.

Group	Method	Bias	RE	SE^{TS}	SE^{ES}	MCSD	CP^{TS}	CP^{ES}
I	un	0.002	1.000	0.106	–	0.105	0.950	–
II	mixed-linear	0.001	1.035	0.104	–	0.104	0.952	–
III	semi-all	0.001	0.990	0.176	0.182	0.106	0.995	0.996
IV	semi-forw- α	–0.001	1.132	0.094	0.096	0.099	0.934	0.944
	semi-forw-aic	–0.001	1.129	0.094	0.096	0.099	0.933	0.941
	semi-forw-bic	–0.001	1.137	0.094	0.097	0.099	0.934	0.944
	semi-ffsr	–0.001	1.134	0.094	0.097	0.099	0.935	0.945
V	semi-scad	–0.001	1.138	0.094	0.097	0.099	0.935	0.942
	semi-lasso	–0.001	1.138	0.094	0.096	0.099	0.935	0.942
	semi-adalasso	–0.001	1.137	0.094	0.097	0.099	0.935	0.944
	semi-ose- $L_{0.01}$	–0.001	1.137	0.094	0.097	0.099	0.935	0.944
	semi-ose-scad	–0.001	1.138	0.094	0.097	0.099	0.935	0.942
benchmark	semi-oracle	0.000	1.198	0.095	0.098	0.096	0.947	0.956

Table 6

Results for the ACTG 175 data. Estimate, estimate of the treatment effect β_1 ; SE, standard error, category I and II using usual formula, group III-IV using the empirical sandwich method; RE, square of SE for the unadjusted method divided by that of the indicated method.

Group	Method	Estimate	SE	RE
I	un	46.81	6.76	1.00
II	an1-all	51.66	5.67	1.42
	an1-missub	46.73	7.16	0.89
	an1-forw- α	51.07	5.55	1.48
	an2-all	50.95	11.24	0.36
	an2-missub	47.07	7.16	0.89
	an2-forw- α	49.75	5.56	1.48
			SE^{ES}	
III	semi-all	56.06	5.13	1.74
IV	semi-forw- α	51.14	5.10	1.75
	semi-forw-aic	53.11	5.03	1.81
	semi-forw-bic	51.34	5.08	1.77
	semi-ffsr	51.32	5.08	1.77
V	semi-scad	53.74	5.06	1.79
	semi-lasso	49.58	5.16	1.72
	semi-adalasso	49.64	5.10	1.76
	semi-ose- $L_{0.01}$	49.64	5.10	1.76
	semi-ose-scad	54.51	5.07	1.78